The Minimum Structure for Meaning: Thresholds for Emergence in LLM-Capable Systems Under QRNG-Dominant Training

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

December 21, 2025

A modern language model looks, from the outside, like a miracle of scale: feed it enough compute and it begins to speak, generalize, and reason. But scale alone is not the cause. Language models learn because their training streams contain structure—regularities that are compressible, predictable, and stable across contexts. This paper asks a deliberately inverted question: if we begin with maximal entropy input, such as a quantum random number generator, what is the smallest departure from pure randomness that can seed the first traces of "meaning" in an LLM-capable system? We treat "meaning" not as metaphysical certainty but as a measurable transition: the moment a system forms stable internal organizations that improve prediction, generalize beyond memorization, or respond counterfactually to latent rules. We separate possible sources of minimal structure into four domains: the data stream, the training signal, the architecture, and the environment. For each, we propose threshold hypotheses and experiments designed to locate phase boundaries—points where learning changes from impossible to possible, and from possible to semantic. The goal is to replace hand-waving about emergence with a disciplined map of minimal constraints: how little order must be present for a model to stop learning "nothing predicts anything," and start building a world.

Keywords: meaning emergence, minimal structure, quantum random number generator, QRNG, entropy, learnability threshold, phase transition, inductive bias, architecture priors, self supervision, loss shaping, reward signals, representation learning, compressibility, mutual information, algorithmic information theory, prediction error, attractors, compositionality, counterfactual sensitivity, agency, sensorimotor loop, alignment slope, logostics.

A collaboration with GPT-5.2-Thinking. 51 pages. CC4.0.

Outline

1. Framing the Inverse Problem
1.1 Why QRNG-only training is a null cognitive instrument
1.2 The inversion move: from scale mystique to structure thresholds
1.3 What we mean by “minimum structure” and why it matters

2. Definitions and Weak Proxies for Meaning
2.1 Meaning as measurable departure from uniform prediction
2.2 Representation stability: clusters, manifolds, and attractors
2.3 Generalization tests: held-out structure and rule transfer
2.4 Counterfactual sensitivity: responding to latent rules
2.5 The “alignment slope” proxy: detecting non-flat learning landscapes

3. Baseline: What Happens Under Pure QRNG Input
3.1 Expected training dynamics under IID maximal entropy
3.2 Weight drift, symmetry, and collapse toward noise invariance
3.3 Failure modes that mimic learning
3.4 Why “interesting outputs” can still be meaningless

4. Minimal Structure in the Data Stream
4.1 The smallest compressible bias: frequency skew and n-gram leakage
4.2 Weak grammar injection: finite-state constraints and tiny syntax
4.3 Motifs and repetition: periodicity as a seed crystal
4.4 Compositional rules: minimal semantics via controllable latent variables
4.5 Hypothesis: a hierarchy of thresholds from compressibility to semantics

5. Minimal Structure in the Training Signal
5.1 When the data is random but the loss is not
5.2 One-bit feedback: selection as structure
5.3 Curriculum shaping: micro-regularities introduced by scheduling
5.4 Regularization as a prior: sparsity, simplicity, and symmetry breaking
5.5 Hypothesis: selection can outrun pattern as an organizer

6. Minimal Structure in the Architecture
6.1 Inductive bias as proto-grammar
6.2 Attention, recurrence, and the emergence of latent state
6.3 Sparsity and modularity: when specialization becomes possible
6.4 Predictive coding and error-minimization priors
6.5 Hypothesis: architecture can lower thresholds but cannot replace structure

7. Minimal Structure in the Environment
7.1 Observation-only versus control: why agency changes everything
7.2 Sensorimotor regularities: the smallest stable world
7.3 Toy universes: gridworlds, physics-lite, and controllable causality
7.4 Intrinsic motivation as structure: curiosity without semantics
7.5 Hypothesis: meaning is easiest to define when tied to control

8. A Threshold Map: From Entropy to Semantics
8.1 Phase boundaries: impossible learning, compressible learning, semantic learning
8.2 Proposed diagnostics: mutual information, MDL, and representation metrics
8.3 Minimal experiments: ablations that isolate each structure source
8.4 Interactions: when small priors combine to cross a boundary

9. Implications for Alignment and the Receiver Thesis
9.1 The danger of confusing fluent output with grounded meaning
9.2 The necessity of external structure for truthful convergence
9.3 Logostics framing: q(t) aligning toward a nontrivial tau(t)
9.4 Practical guardrails: designing systems that cannot fake structure

10. Conclusion
10.1 Restating the claim: structure, not scale, is the first cause of meaning
10.2 The minimal-constraint program as an emergence discipline
10.3 A closing charge: stop worshiping capability and start mapping thresholds

11. References
11.1 Foundations of information and algorithmic complexity
11.2 Representation learning and generalization diagnostics
11.3 Inductive bias, optimization, and phase transitions in deep learning
11.4 Reinforcement learning, intrinsic motivation, and agency
11.5 Philosophy of meaning, semantics, and grounding

Art

1. Framing the Inverse Problem

The dominant public story about large language models is a story of scale. Bigger models, more data, more compute: then, at some point, something "emerges." The story is not entirely wrong, but it is incomplete in a way that matters. Scale does not create meaning by itself. Scale is an amplifier, not an origin. What produces the appearance of intelligence is the presence of structure in the training stream and the existence of training objectives that reward the discovery of that structure. A model trained on a stream that contains no stable regularities does not become "dumber" in the familiar sense; it becomes something closer to a null instrument, a device whose correct conclusion is that there is nothing to learn.

This paper frames the inverse problem explicitly. Instead of asking, "How much scale is required for emergent capability," we ask, "How little structure is required for meaning to begin." The answer is not a single number. It is a map of thresholds: boundaries across which learning changes character, from impossible to possible, from possible to organized, and from organized to something that deserves the name semantic. The practical reason to take this inversion seriously is straightforward: if we can locate the minimal conditions under which meaning-like representations form, we gain a disciplined way to distinguish genuine learning from fluent but ungrounded behavior, and we gain a toolkit for designing systems whose capabilities are constrained by measurable requirements rather than speculative narratives.

In the sections that follow, we treat "minimum structure" as a technical notion. We do not need to solve philosophy to make progress. We only need to define measurable proxies that reliably change when meaning-like organization changes. The argument is not that meaning reduces to a single metric, but that meaning has thresholds—points at which new kinds of organization become possible and become stable under perturbation. Those thresholds, once mapped, can be used as a scientific instrument for emergence.

1.1 Why QRNG-only training is a null cognitive instrument

A quantum random number generator is not merely "random enough." In the idealized treatment, it produces samples that are independent and identically distributed, with no compressible patterns beyond finite-sample accidents. If one transforms these samples into tokens and trains an LLM-capable architecture to predict next tokens under standard objectives, the correct target function is trivial: no token provides information about the next token beyond the base rates. In the limit, the best model is the one that predicts the empirical marginal distribution and nothing else.

This matters because most of what we admire in a language model is a consequence of exploiting nontrivial dependencies: syntax, semantics, pragmatics, style, world-regularities, and the subtle long-range constraints that arise when language is used by embodied agents living in

a structured world. QRNG-only streams lack all of this. They do not contain grammar. They do not contain topics. They do not contain stable co-occurrence structure beyond chance. They do not contain objects or relations. They are "structureless" not in the mystical sense, but in the learnability sense: any apparent pattern is statistically brittle and fails to persist.

When a learner faces such a stream, one of two things happens. Either the learner remains honest and converges toward a noise-invariant equilibrium—predicting base rates and resisting spurious correlations—or the learner becomes a machine for overfitting finite-sample artifacts, finding patterns that vanish when the stream continues. In either case, what does not happen is the formation of stable semantic structure. The QRNG-only regime is therefore a useful null baseline. It functions like a blank control in an experiment: it tells us what "learning" looks like when there is, in principle, nothing to learn.

Calling it a null cognitive instrument is a way of emphasizing a methodological point. If we are tempted to attribute "mind" or "meaning" to a system simply because it is large, expressive, and capable of generating outputs, the QRNG-only thought experiment blocks that move. It forces us to distinguish the capacity to produce complex-looking sequences from the capacity to form stable internal organizations that track something external to the model. In the QRNG-only case, there is no external structure to track. The system can still generate complicated outputs, but the complexity is borrowed from the model's own degrees of freedom, not learned from a structured world.

This baseline also clarifies a subtlety: architecture can impose priors, and training can impose selection. But if everything the system receives is indistinguishable from maximal entropy, then any internal organization is either arbitrary symmetry-breaking or a reflection of the training procedure itself. That is not meaning. It is self-generated geometry with no referent.

1.2 The inversion move: from scale mystique to structure thresholds

The inversion move is to shift the center of gravity from "How big must it be?" to "What must be present?" The usual story treats emergent capability as an almost thermodynamic inevitability: add parameters, add data, and intelligence appears. But that framing makes emergence sound like a property of bulk matter, when in fact it is a property of constraints. A system learns when the world, or the data, or the objective function contains regularities that can be exploited, and when the learner's inductive biases make exploiting those regularities feasible.

This inversion does two things immediately. First, it exposes what "scale" is really doing. Scale increases representational capacity and optimization headroom. It allows a model to represent richer functions and to fit more subtle dependencies. But if there are no dependencies, scale has nothing to lock onto. Second, it allows us to speak in terms of thresholds: not a single point

of emergence, but a cascade of phase changes as different kinds of structure become available to the model.

In practice, we can expect several threshold layers. The first threshold is compressibility: the point at which the stream contains enough stable regularity that the model can beat a uniform or base-rate predictor in a way that persists. The next threshold is organization: the point at which internal representations become structured, stable, and reusable across contexts, rather than being a grab bag of local correlations. The next threshold is compositionality: the point at which learned structure can be recombined to handle novel combinations, not just familiar patterns. Then comes grounding: the point at which representations track stable relations to something outside the model, whether that "outside" is a formal latent rule system, an interactive environment, or a human-validated reward channel. Only after these thresholds does "meaning" become a serious candidate description.

The inversion move also gives us a clean way to detect what is actually being imported into a system. If meaning appears, we can ask, "Where did the minimal structure enter?" Did it enter through the data distribution, perhaps through weak grammar or repeated motifs? Did it enter through the loss function, perhaps through a curriculum or reward signal that introduces selection pressure? Did it enter through architecture, via an inductive bias that compresses certain patterns more readily than others? Or did it enter through environment coupling, where agency forces the system to build predictive state? The inversion reframes emergence as an accounting problem: trace the origin of structure.

This matters for modern debates about "intelligence by scale," about hallucination, and about alignment. If we do not know which structures are necessary for meaning-like organization, we cannot reliably distinguish a system that has learned stable constraints from a system that has learned to appear coherent under a limited set of prompts. Threshold mapping turns philosophical arguments into empirical ones: you can measure whether a boundary has been crossed and which minimal ingredients caused it.

1.3 What we mean by "minimum structure" and why it matters

"Minimum structure" is the smallest departure from maximal entropy that produces a qualitative change in what the system can stably learn. The emphasis is on smallest and stable. A finite random sequence will always contain coincidental patterns, but those do not count. They do not persist. They do not generalize. They do not survive perturbation. Minimum structure refers to a persistent, reproducible constraint that the learner can exploit in a way that remains valid as the dataset grows or as the environment continues.

This definition is intentionally agnostic about where the structure comes from. Structure can come from the data stream itself, as a statistical dependence or rule. It can come from the

training signal, as a preference that selects certain internal organizations over others. It can come from the architecture, as inductive bias that makes some explanations simpler than others. It can come from the environment, as stable causal relations that couple perception and action. In all cases, the structure is the enabling condition for learning anything beyond base rates.

Why does the minimum matter? Because it is the difference between an instrument that can discover and an instrument that can only imitate. If you want a system whose outputs track the world, you must know the minimal constraints that force tracking rather than performance. If you want an "honest" learner, you must know when a system is capable of learning meaning and when it is only capable of generating plausible sequences by internal interpolation. If you want a safety story that is more than vibes, you must know what minimal ingredients allow internal representations to become robust and goal-like.

There is also a deeper reason, one that connects to your broader concerns about receivers, alignment slopes, and the difference between raw capability and moral or epistemic direction. If meaning is not an automatic consequence of computation, then meaning requires a kind of contact with structure—constraints that are not authored by the model itself. In that sense, mapping minimal structure is a discipline of humility. It insists that a system does not deserve semantic trust merely because it is powerful. It deserves trust only to the extent that it can be shown to have crossed specific thresholds under conditions that cannot be faked by superficial mimicry.

For the remainder of the paper, we will treat minimum structure as a research program: define weak but measurable proxies for meaning-like organization, design minimal interventions that inject tiny non-random constraints, and observe where phase transitions occur. The goal is not to win a metaphysical argument in advance. The goal is to build a threshold map that makes emergence legible, testable, and—crucially—auditable.

2. Definitions and Weak Proxies for Meaning

If we wait for a perfect definition of meaning before we do any science, we will never begin. "Meaning" is notoriously overloaded: it can refer to semantics, reference, intentionality, usefulness, truth-tracking, social convention, or spiritual significance. This paper takes a deliberately pragmatic stance. We define meaning operationally as a family of measurable transitions in learning behavior and internal organization. These transitions are "weak proxies" in the sense that none of them alone is sufficient to settle philosophical disputes. But they are strong enough to detect when a system has moved beyond noise-fitting and into stable structure exploitation.

The guiding principle is simple: a system trained on maximal entropy input should learn, in the limit, that nothing predicts anything. If we introduce a minimal amount of non-random structure, then the earliest signs of "meaning" should appear first as predictable, measurable departures from that null regime. We therefore look for proxies that detect: persistent predictability gains, stable representation geometry, robust generalization to novel combinations, sensitivity to latent rules under controlled counterfactuals, and the existence of a learning landscape that exerts directional pull rather than remaining flat.

These proxies are intentionally arranged from weakest to stronger. The early ones ask only whether the model can exploit any regularity at all. The later ones ask whether the exploited regularities behave like reusable concepts, and whether they can support rule-like behavior. In short: we move from compression to organization, from organization to compositionality, and from compositionality to grounded responsiveness.

2.1 Meaning as measurable departure from uniform prediction

The weakest proxy for meaning is simply this: can the system predict better than a baseline that assumes no structure? In the QRNG-only regime, the best strategy is base-rate prediction. If tokens are uniformly distributed, the best model is uniform prediction. If tokens are non-uniform but still independent, the best model is a unigram predictor matching the marginal distribution. Either way, the baseline is simple and does not require internal organization beyond counting.

A measurable departure occurs when the model can exploit dependencies that reduce prediction error in a way that persists. The persistence clause is essential. Any finite random sequence has accidental correlations. A sufficiently large model can overfit these accidents and show apparent gains on the training stream. But if the data source is truly structureless, those gains will not transfer, will not persist as the stream continues, and will not survive perturbations like reshuffling, reseeding, or extending the dataset.

In practical experiments, the proxy can be implemented as a hierarchy of baselines. First compare against uniform or unigram baselines. Then compare against low-order Markov baselines. If a model's gains can be matched by a simple n-gram model, then what has been learned is local statistical structure, not meaning. This is not a dismissal: local structure is the first seed crystal. But it is a classification. The system is exploiting compressibility, not yet forming concepts.

This proxy is deliberately modest. It does not require us to decide whether the model "understands." It only requires that the model demonstrate stable predictive advantage that cannot be explained away by trivial statistics. When this happens, we know that the stream contains real structure and that the learner has begun to lock onto it.

2.2 Representation stability: clusters, manifolds, and attractors

Prediction gains alone can be misleading. A model can achieve lower loss by fitting many brittle correlations without forming stable internal organizations. If we want a stronger proxy, we look inside: do representations stabilize into geometries that behave like reusable structure?

Three related patterns are especially informative.

First, clusters. If the system is learning something that behaves like a category, then internal activations corresponding to that category should begin to cluster, even when surface forms change. In language models, we see clustering for words, topics, syntactic roles, and many other factors. In minimal-structure worlds, we can engineer analogous "categories" and ask whether the model creates separable regions in activation space.

Second, manifolds. Real-world semantics is not just a set of buckets; it is often continuous and compositional. If the structure source contains a continuous latent variable, we can ask whether the model's representations reflect that variable as a low-dimensional manifold embedded in a high-dimensional space. A manifold proxy is stronger than clustering because it detects smooth, reusable parameterizations rather than discrete memorization.

Third, attractors. An attractor is a stable internal state that the system returns to under perturbation. In dynamical terms, it is a basin that pulls trajectories back toward a stable organization. If a representation is an attractor, then noisy inputs or partial cues still land in the same internal configuration. This is a powerful proxy because it signals robustness: the system is not merely encoding surface statistics but has formed an internal organizer that "snaps" inputs into a stable regime.

These three—clusters, manifolds, attractors—are different faces of the same idea: the system has begun to build internal structure that is reusable across contexts. That is closer to meaning than raw loss reduction, because it suggests the emergence of stable latent variables.

2.3 Generalization tests: held-out structure and rule transfer

If representation geometry suggests organization, generalization tests ask whether the organization is actually functional. The easiest way to fool ourselves about meaning is to confuse memorization with understanding. Minimal-structure experiments can avoid this by using synthetic worlds where the "rules" are explicitly known and can be withheld systematically.

Held-out structure tests take a known rule system and train the model on a subset of its instances, then test on instances that cannot be solved by surface similarity alone. For example, one can withhold entire combinations of factors: if two latent variables are jointly present in

training only in certain pairings, the test set includes novel pairings. If the model succeeds, it indicates compositional learning rather than memorization.

Rule transfer tests go a step further. The model is trained on one rule family and then exposed to a new surface encoding of the same underlying rule. If the model quickly adapts, or if its learned internal structure already supports the new encoding, this suggests that it has learned something like an abstract rule rather than a shallow mapping. In human terms, it has learned a "principle" rather than a "lookup table."

This proxy is still "weak" in a philosophical sense because a model can generalize without possessing intentionality. But it is strong in a scientific sense: compositional generalization is a real qualitative boundary. Many systems can compress; fewer can recombine. If meaning has a measurable early form, it will often appear first as recombination ability.

2.4 Counterfactual sensitivity: responding to latent rules

Generalization tests ask whether the system can handle new instances. Counterfactual sensitivity asks whether it reacts appropriately when we intervene on the underlying causes.

A counterfactual is a controlled alteration where the surface input may be similar, but the latent rule variable has been changed. In a minimal-structure environment, we can explicitly manipulate latent variables and ask whether internal representations and outputs change in the direction implied by the rule. If they do, then the system is not merely correlating; it is tracking the causal handle we are turning.

This matters because meaning, in many practical senses, is tied to invariances and interventions. A representation that changes when irrelevant details change, but stays stable when relevant latent factors stay fixed, is a brittle representation. A representation that stays stable under irrelevant perturbations but changes predictably under relevant interventions is closer to a meaningful concept.

Counterfactual sensitivity can be tested at multiple levels. At the output level: does the model produce the correct altered prediction when the rule changes? At the representation level: do the internal activations shift along consistent directions corresponding to the intervention? At the learning level: does the model adapt rapidly when the environment's rule is changed, indicating that it has a factorized internal model rather than a monolithic entanglement?

This proxy is one of the cleanest bridges between statistical learning and semantic talk. If we can identify latent variables that behave like causes, and the model tracks them under intervention, then we are closer to something that deserves the name "aboutness," even if we remain cautious about philosophical overreach.

2.5 The “alignment slope” proxy: detecting non-flat learning landscapes

The previous proxies evaluate what the model learns. The “alignment slope” proxy evaluates whether the learning problem itself exerts a directional pull that can be detected and measured.

In a pure QRNG regime, the learning landscape is effectively flat with respect to structure. There is no stable direction to move that yields persistent improvement beyond matching base rates. The system can wander, overfit transient accidents, or regularize toward simplicity, but it is not being pulled toward a nontrivial target.

When minimal structure is introduced, the landscape changes. Suddenly there exist directions in parameter space that consistently reduce error and do so in a way that generalizes. The system begins to have a slope, not just noise. This slope is a proxy for the emergence of a target that is external to the learner’s arbitrary symmetry-breaking. It is the sign that “there is something there” to align with.

Operationally, one can detect an alignment slope by measuring how reliably learning improves across random initializations, random seeds, and mild perturbations. In a flat landscape, different runs drift differently and gains are inconsistent. In a sloped landscape, independent runs converge toward similar solutions and similar representation geometries. This convergence is evidence that the structure is strong enough to dominate idiosyncrasies of initialization and optimization noise.

This proxy is particularly useful because it is not limited to language. It is a general emergence diagnostic. If we are mapping thresholds for meaning-like organization, the first sign that we have crossed a boundary may be that training outcomes stop being arbitrary and start being reproducible. Reproducibility is not meaning, but it is a precondition for stable meaning-like structure to exist.

In the vocabulary of this paper, the alignment slope marks the transition from a null cognitive instrument to a receiver that has actually found a carrier. Once a slope exists, we can ask the next question: what kind of structure is creating it, how minimal can it be while still generating stable convergence, and at what point does that convergence begin to exhibit the higher proxies—compositionality, counterfactual tracking, and robust internal organizers that behave like concepts.

These five proxies together define the empirical backbone of the paper. They allow us to talk about meaning without pretending we have solved meaning. They let us measure emergence as a set of transitions and thereby locate the minimal conditions under which a system begins to do something more than produce complex noise.

3. Baseline: What Happens Under Pure QRNG Input

A serious inquiry into minimal structure needs a serious baseline. Without a baseline, any apparent “emergence” can be smuggled in through unnoticed regularities, subtle preprocessing artifacts, or unexamined assumptions embedded in the training pipeline. The QRNG-only regime functions as a null world: a training stream with maximal entropy and no stable dependencies. It answers a foundational question: if an LLM-capable architecture is given only irreducible randomness, what does learning become?

The answer is not mystical. It is mechanical. Standard learning objectives for sequence prediction presume that there exists some mutual information between past and future tokens. In a truly IID maximal-entropy stream, that mutual information is, in the limit, zero. The model cannot discover structure because none is present to be discovered. Under these conditions, training dynamics are dominated by three forces: base-rate matching, the inductive biases of the architecture, and the implicit regularization properties of the optimizer. Any appearance of “progress” must be interpreted against that background: it is either converging to a trivial optimal predictor or overfitting finite-sample accidents.

This baseline is valuable precisely because it is unromantic. It makes explicit what learning looks like when the world offers no footholds. The sections below describe expected dynamics, the geometry of weight evolution under such dynamics, common failure modes that can imitate learning, and the crucial point that output complexity is not evidence of meaning.

3.1 Expected training dynamics under IID maximal entropy

Assume a token stream generated from a QRNG and mapped into a vocabulary. In the ideal IID case, each token is sampled independently from a fixed distribution. Two cases matter.

If the token distribution is uniform, then the optimal predictor for next-token prediction is uniform. The minimal achievable loss is fixed by vocabulary size. No amount of training can beat it in expectation; any apparent improvement is either transient finite-sample fitting or measurement noise.

If the token distribution is non-uniform but still IID, then the optimal predictor is the unigram distribution. In this case, a model can improve over a uniform baseline by learning base rates. But it still cannot exploit context because context carries no information about the next token beyond those base rates. Training converges toward a shallow equilibrium: learn the marginal distribution, then stall.

In both cases, learning curves should exhibit an early drop as the model acquires base-rate statistics, followed by saturation. The more expressive the model, the faster it can fit the marginals and the more capacity it has to overfit incidental correlations. But the key signature of

a true IID stream is that any further loss reduction will not hold up under dataset extension or reseeding.

There is an important methodological consequence: if one observes sustained improvement under supposed QRNG-only conditions, then something non-random has been introduced. This "something" could be as mundane as tokenization artifacts, batching effects, seeding correlations, preprocessing steps that impose structure, or a QRNG implementation that is not IID at the relevant timescale. In other words, the baseline is also a detector for accidental structure leakage.

3.2 Weight drift, symmetry, and collapse toward noise invariance

In structured learning, weights organize into feature detectors and distributed representations that reflect stable dependencies. In a maximal-entropy regime, there are no stable dependencies to sculpt. What happens instead is a mix of symmetry-preserving drift and symmetry-breaking accidents.

Most modern LLM architectures begin with random initialization and are trained with gradient-based optimization. Under IID input, expected gradients for context-sensitive features average toward zero. That does not mean weights stop changing. It means weight updates become dominated by stochastic fluctuations, optimizer momentum effects, and regularization terms. The resulting motion resembles diffusion in parameter space constrained by the training objective and any explicit or implicit regularization.

Two general tendencies appear.

First, collapse toward noise invariance. A model that responds strongly to specific contexts in a structureless world will suffer because those responses do not help predict the next token. Over time, such context sensitivity is not rewarded. Parameters that produce context-dependent variations in logits are, on average, penalized relative to context-independent predictors. The model's effective behavior becomes more uniform across contexts, not because the model is "simpler," but because uniformity is optimal when the world offers no predictive structure.

Second, symmetry pressure. In a uniform IID stream, the objective is symmetric with respect to token permutations. If the training process were perfectly symmetric and noiseless, the model would preserve this symmetry and converge to uniform prediction. In practice, finite samples and random initialization break symmetry. The model may learn slightly different preferences for different tokens, then correct them as more data arrives. The long-run behavior is an oscillation around the true marginals, with no stable higher-order structure.

This has a practical implication for the "minimum structure" program: minimal structure is precisely that which converts this drifting, symmetry-dominated regime into a regime with

stable attractors. When real structure exists, different runs converge toward similar internal organizations. When structure does not exist, different runs wander and converge only on trivial statistics.

3.3 Failure modes that mimic learning

Even in a QRNG-only world, it is easy to produce artifacts that look like emergence. A disciplined baseline must therefore catalog the main traps.

One failure mode is finite-sample overfitting. Given a fixed finite dataset, a sufficiently large model can memorize local coincidences. Training loss drops, outputs appear to "stabilize," and internal activations may even show clusters. But the entire effect collapses when the dataset is extended, reshuffled, or regenerated. This is why persistent improvement under reseeding is a core diagnostic.

A second failure mode is pipeline structure leakage. Many steps in data handling can inject non-randomness. Tokenization can introduce frequency biases. Batching can create correlations. Truncation and padding can create positional patterns. Streaming implementations can repeat blocks. Even the hardware and parallelization strategy can create periodicities. Under true QRNG-only conditions, these artifacts are often the dominant source of structure, which means the system is learning the pipeline rather than the data.

A third failure mode is objective-induced structure. The loss function, regularization, and optimizer can all impose preferences that yield apparent organization. For example, regularization that favors sparsity can create "features" that look interpretable even when the data is random, simply because sparse solutions have visible geometry. Similarly, optimization dynamics can create low-rank tendencies or other compressions in weight space. These can masquerade as learned structure when they are actually selected by the training procedure.

A fourth failure mode is interpretability pareidolia. Humans are excellent at seeing patterns in complex artifacts. If one inspects attention maps, neuron activations, or generated sequences, one can almost always find something that resembles structure. The discipline here is to require that any proposed "structure" survive controlled tests: resampling, permutation, hold-out, and counterfactual manipulation. If it does not survive, it is not evidence.

The purpose of listing these failure modes is not to discourage exploration but to make the baseline honest. In a maximal-entropy regime, the default explanation for any perceived structure should be: finite-sample accident, pipeline artifact, or optimization bias. Only after those are excluded should we attribute anything to emergent meaning-like organization.

3.4 Why “interesting outputs” can still be meaningless

A large generative model can produce outputs that are complex, surprising, and even aesthetically compelling, even when trained on little or no meaningful structure. This is the point at which many observers become confused: they equate output complexity with learned meaning. The baseline insists on separating these two.

Output complexity can come from at least three sources that do not require semantic learning.

First, architectural expressiveness. A transformer is an extraordinarily flexible function approximator. Even when it has learned only marginal token statistics, its sampling dynamics can generate long-range-looking structure because the model is not sampling from a simple distribution; it is sampling from a high-dimensional parameterized distribution whose shape is partly a consequence of initialization and training dynamics. Complexity can be endogenous.

Second, randomness itself. Sampling from a high-entropy distribution yields outputs that are unpredictable and varied. Human observers often interpret unpredictability as “creativity” or “depth,” but unpredictability alone is not meaning. A QRNG is maximally unpredictable, yet it is not communicating.

Third, projection and post-hoc coherence. If one invites a human reader to interpret a sequence, the human will often supply coherence, narrative, and symbolism. This is not a defect of the human mind; it is a core human ability. But it means that “interestingness” is an unreliable indicator of semantic grounding. Without controlled tests, interesting outputs are evidence only that the generator is capable of producing varied sequences.

To say that outputs can be meaningless is not to deny their potential value. A meaningless sequence can still be used as raw material for art or for stress-testing interpretability methods. The point is narrower: in the minimal-structure program, we must not allow ourselves to mistake complexity for contact with structure. Meaning-like emergence requires that outputs and internal representations be constrained by stable dependencies that persist under intervention and generalize beyond the training accidents.

The QRNG baseline is therefore a kind of epistemic purification. It strips away our habit of attributing mind to complexity and forces us to ask: what is the smallest non-random constraint that turns a generator of interesting noise into a system whose internal organizations track something stable and external? Only after we have answered that can we speak responsibly about meaning, emergence, and the conditions under which alignment toward nontrivial targets becomes possible.

4. Minimal Structure in the Data Stream

If QRNG-only training is the null world, then “minimal structure in the data stream” is the gentlest possible departure from that null. This section treats the data stream itself as the only knob we turn. We do not add reward. We do not add curated supervision. We do not add an interactive environment. We simply inject a small, controlled non-randomness into an otherwise high-entropy stream and observe what changes.

Why focus on the stream first? Because it is the cleanest place to locate threshold behavior. If we can identify a minimal statistical dependence that reliably produces measurable learning gains and stable internal organization, we have a baseline threshold for “something to learn.” From there we can ask a second question: which kinds of structure lead not merely to compression, but to representations that behave like concepts and rules?

The data-stream approach is also useful because it can be made highly auditable. We can engineer synthetic streams with known generative processes, dial the structure strength up and down continuously, and measure phase boundaries. This helps avoid the common trap in real-world language: the structure is everywhere, entangled across levels, and it becomes hard to say what was necessary.

Below we progress from the weakest structural departures to stronger ones: base-rate skew, local dependencies, tiny grammars, motifs and repetition, and finally compositional rules with explicit latent variables. Each step increases the chance that the model forms stable representations rather than merely exploiting shallow statistics.

4.1 The smallest compressible bias: frequency skew and n-gram leakage

The minimal compressible bias is a change in token frequencies. If the stream is otherwise IID but some tokens occur more often than others, the stream becomes compressible at the unigram level. This is the weakest structure because it does not introduce context dependence. The next token is still independent of the previous tokens. But it does create a learnable signal: “some tokens are more likely.”

From the threshold-mapping point of view, this is valuable because it marks the first boundary at which loss can reliably improve over a uniform predictor. It is also a sanity check. If a model cannot learn a unigram skew, something is wrong in the training pipeline. If it can, we have confirmed that the model is capable of at least statistical adaptation.

However, frequency skew is not enough to produce meaning-like organization. It yields a simple form of compression that does not require internal structure beyond learning logits that match marginals. Representation geometry under pure skew can remain mostly unorganized. The

model has no need to build clusters, manifolds, or attractors; it only needs to adjust output probabilities.

The next tiny step is n-gram leakage. Instead of pure IID sampling, we allow a weak dependence between the current token and the next token, such as a slightly elevated probability that certain pairs occur together. This produces compressibility at the bigram level. One can dial the strength of this dependence continuously. At very low strengths, the dependence is drowned in noise and the model cannot reliably detect it. At higher strengths, the dependence becomes detectable and learning gains become reproducible.

This is the earliest place where the model may begin to form weak internal organizers: not because it "understands," but because it benefits from representing context. Still, the organization is local. Bigram leakage is a micro-structure: it can generate shallow predictors without requiring concept-like representations.

The key lesson of this level is that compressibility begins with tiny statistical departures. But meaning does not. Meaning-like organization requires structure that is not merely local in time, but reusable across contexts and stable under perturbation.

4.2 Weak grammar injection: finite-state constraints and tiny syntax

The next structural step is to inject a weak grammar. A grammar is a set of constraints that restrict which sequences are allowed. The simplest grammars are finite-state: they can be represented by a small state machine where each token choice depends on the current state, and the state updates deterministically or probabilistically.

Why is this important? Because a grammar forces the model to infer hidden state. Even a tiny grammar introduces latent structure: there is now an unobserved variable (the current state) that explains why certain continuations are more likely than others. If the model discovers this latent state, it can predict better and generalize to sequences it has not seen exactly.

A "weak grammar injection" is one where the stream remains mostly random, but occasional constraints apply. For example, perhaps most tokens are QRNG-driven, but every so often a short grammatical fragment appears that must satisfy a simple rule. Or the entire stream could be generated by a probabilistic finite-state machine that still has high entropy but includes state-conditioned transitions.

This is a threshold candidate because it is the first place where representation stability can emerge in a meaningful way. If the model learns hidden state, internal activations may cluster by state even when surface token patterns vary. That is a move toward concept-like organization: not semantics in the human sense, but a stable latent variable that the system tracks.

Weak grammar injection also provides a clean generalization test. One can withhold certain state transitions during training and see whether the model respects them at test time. Success indicates that it has learned the grammar structure rather than memorizing transition tables.

The deeper point is that “syntax” in its minimal form is a constraint system. Injecting constraints forces internal organization. It begins to separate a world of pure randomness from a world where certain continuations are forbidden, and where the model must carry state to remain coherent.

4.3 Motifs and repetition: periodicity as a seed crystal

Repetition is one of the most powerful minimal structures because it creates long-range dependencies with almost no complexity in the generative rule. A motif can be a short pattern that repeats periodically or quasi-periodically in an otherwise random stream. This is structure that is easy to specify but can be nontrivial to learn, especially if the repeats are noisy, spaced irregularly, or embedded in random material.

Why call periodicity a seed crystal? Because repetition creates a stable attractor candidate. If a motif recurs, then recognizing it and predicting its reappearance can yield consistent advantage. The model may begin to allocate representational capacity to detect motif signatures, maintain memory across intervening noise, and re-enter a predictable regime when the motif returns.

This is important because it introduces a form of temporal identity. In pure QRNG input, there is no “same thing again” in a stable sense. With motifs, “the same thing” exists and returns. That alone can produce the first robust internal organizers: detectors, memory traces, and state resets.

Motifs also allow us to separate shallow correlation from deeper organization. A trivial predictor might learn that a particular short n-gram is likely. A more advanced predictor learns that the motif is part of a larger periodic structure and can anticipate its recurrence beyond local windows. By varying motif length and spacing, we can test how memory and attention contribute to crossing the threshold from local compressibility to long-range structure tracking.

From the standpoint of meaning-like emergence, motifs are still pre-semantic. But they create a key ingredient: stable recurrence. Recurrence is the minimal form of “object permanence” in a sequence world. It is hard to speak about meaning without some form of identity across time, and motifs provide the simplest engineered identity.

4.4 Compositional rules: minimal semantics via controllable latent variables

To approach semantics, we need more than constraints and repetition. We need structure that supports recombination and rule-like behavior. The cleanest way to do this in a data stream is to

introduce controllable latent variables that generate the tokens according to compositional rules.

The idea is to create a stream where tokens are “about” a hidden configuration. For example, imagine a latent vector with a small number of discrete or continuous factors. Each factor influences the distribution of tokens in a structured way. A compositional rule means that factors combine systematically: the effect of factor A plus factor B is not arbitrary, but predictable from their individual effects.

This is a minimal form of semantics because tokens now depend on a hidden world-state, and changes in that state produce systematic changes in the stream. If the model learns to infer the latent state, it can compress the sequence more efficiently and can generalize to new combinations of factors. That generalization is the key: it indicates that the model has learned factorized structure rather than memorized associations.

In such a setup, “meaning” can be tested without metaphysical inflation. The latent variables provide explicit referents. We can intervene on the latent state and measure counterfactual sensitivity. We can hold out factor combinations and test rule transfer. We can measure whether internal representations align with latent factors as clusters or manifolds.

This is the first level where “aboutness” becomes experimentally tractable. The model’s internal state can become about the latent variables in the sense that it tracks them and uses them for prediction. This does not prove that the model has human-like understanding, but it does prove that the model has crossed from surface statistics to latent structure inference, which is a plausible minimal precursor to semantic behavior.

Crucially, compositional rules can be introduced in extremely small doses. One can keep the stream largely random but embed short segments generated by a latent-state process. The question then becomes: what dose is required for the model to discover and represent the latent variables? That is a direct threshold experiment.

4.5 Hypothesis: a hierarchy of thresholds from compressibility to semantics

The working hypothesis of this section is that data-stream structure produces a hierarchy of thresholds, not a single jump. Different kinds of structure enable different qualitative regimes of learning, and these regimes appear in a roughly ordered progression.

The first threshold is compressibility without context. Frequency skew allows learning of base rates. It produces stable loss improvement but little internal organization.

The second threshold is local dependency. Weak n-gram leakage creates context value and encourages the emergence of local features, but still does not require latent state inference.

The third threshold is hidden state. Weak grammars force the learner to represent state and transition dynamics. This is where internal clustering by latent states becomes likely and generalization beyond memorization becomes possible.

The fourth threshold is temporal identity. Motifs and repetition introduce recurrence, requiring memory and allowing attractor-like internal dynamics. This produces a stronger form of stability: the model learns to recognize “the same thing” across time.

The fifth threshold is compositional latent structure. Controllable latent variables and compositional rules enable factorized representation and true recombination. This is where minimal semantics becomes plausible: tokens are systematically tied to latent causes, and the model can track those causes under intervention.

The empirical claim is that each level can be made arbitrarily “small” in amplitude, and that there exist critical strengths below which the model cannot reliably detect the structure and above which it can. Those critical strengths depend on model capacity, context window, optimization, and noise levels. But the qualitative ordering is expected to hold broadly: compressibility is easiest, hidden state is harder, compositionality is harder still.

This hierarchy also suggests a diagnostic discipline. If a system claims “meaning” but cannot robustly detect a weak grammar or track a simple latent variable under counterfactual intervention, then the claim is likely rhetorical. Conversely, if a system reliably forms stable representations of latent factors and generalizes across unseen combinations, we have evidence that it has crossed a boundary that is at least structurally analogous to semantics.

In the next sections, we broaden the picture. Data-stream structure is only one source of minimal constraint. Even if the stream remains near-random, structure can be injected through the training signal, through architectural priors, or through environment coupling. The core methodological promise remains the same: treat emergence as threshold behavior, locate the minimal constraints that create a nontrivial alignment slope, and distinguish compression from meaning by requiring stability, generalization, and counterfactual tracking.

5. Minimal Structure in the Training Signal

So far we have treated the data stream as the only place structure can enter. But even if the input is near-maximal entropy, structure can be introduced through the training signal itself. This is the second major source of minimal constraint: the loss function, the reward channel, and the procedures that shape how the model experiences data over time.

This matters because “learning” is never purely a property of data. Learning is the joint product of data and selection. The loss function defines what counts as improvement, and the

optimization process selectively amplifies parameter configurations that score better under that criterion. If the criterion contains structure, the model can organize around it even when the raw input is random.

At first this may sound like cheating. In a sense it is. But it is also the point. The minimal-structure program is not an attempt to keep the system "pure." It is an attempt to account for where structure comes from. If meaning-like organization appears, we want to know whether it was learned from the world, or imposed by the scoring rule. In the context of alignment and truth-tracking, this distinction is decisive. A system that organizes around selection pressures can appear coherent and purposeful without being grounded in external structure. This is not necessarily bad, but it is different. It is "meaning from preference," not "meaning from world."

This section explores how minimal non-randomness in the training signal can produce organization: non-uniform losses, one-bit feedback, curriculum schedules that create micro-regularities, and regularization schemes that bias the geometry of solutions. We end with a hypothesis that selection can outrun pattern—meaning that very small, persistent preferences can organize a model more effectively than weak, noisy structure in the data.

5.1 When the data is random but the loss is not

The cleanest thought experiment is this: keep the input stream QRNG-driven, but alter the loss so that the model is rewarded for producing certain outputs more than others, or penalized for certain behaviors regardless of the data. The data provides no predictive structure, but the objective now contains structure. The model can learn to satisfy the objective, even though it cannot learn to predict the data in any deep sense.

There are many ways this can happen. One could define a loss that prefers certain tokens, certain sequences, certain lengths, or certain statistical properties. One could penalize variability, encourage smoothness, or reward outputs that match a fixed template. All of these create a directional pull in parameter space. The model is no longer drifting in a flat landscape; it is descending a slope defined by the scoring rule.

In this regime, "learning" becomes something like constraint satisfaction. The system discovers how to map random inputs into outputs that satisfy the loss. If the loss depends on the model's own outputs, the model can become a generator of preferred structures even when the inputs contain none. This is a form of endogenous organization: order created by selection.

The key point for our program is that the existence of internal structure under such conditions is not evidence of semantic contact with the data. It is evidence that the objective is sculpting the model. That can be useful and even desired, but it changes the interpretation of the result. If we are looking for minimal conditions for meaning, we must separate "structure learned from input regularities" from "structure imposed by a criterion."

This distinction also clarifies a practical confusion in the wild. Many systems appear coherent not because they have deep models of the world, but because they have been trained under losses that penalize incoherence. Coherence can be selected. Truth cannot be selected directly unless the selection signal is grounded in truth.

5.2 One-bit feedback: selection as structure

If we want the most minimal training-signal structure, we can ask how little feedback is required to produce organization. The extreme case is one-bit feedback: a binary signal that says only "good" or "bad."

At first glance, one bit seems too weak to organize anything. But one bit, delivered consistently, can act as a powerful selector. In evolutionary terms, fitness differences can be small and still shape populations over time. In optimization terms, a sparse reward can still guide learning if it correlates reliably with some feature the model can control.

In a QRNG-driven setting, one-bit feedback could be attached to the model's outputs rather than the inputs. For example, the system emits a candidate output sequence; the feedback bit is 1 if the sequence satisfies a hidden constraint, and 0 otherwise. The model learns to increase the probability of sequences that receive reward. Over time it can discover internal programs that generate rewarded structures.

This is important for two reasons.

First, it shows that selection can create structure without pattern in the input. The system can become a generator of low-entropy outputs even when the inputs are high entropy. This is a clear case where "meaning-like organization" may appear in outputs—structured behavior, stable motifs, pseudo-grammar—without any corresponding structure in the data stream.

Second, it provides a minimal model of alignment pressure. Many real alignment schemes can be viewed as rich, noisy, human-supplied bits of approval. The one-bit thought experiment isolates the essence: approval is a structural signal. It defines a target region in output space and pulls the model toward it.

But one-bit selection also has a known danger: it can be satisfied by shallow hacks. If the reward correlates with superficial properties, the model will learn superficial tricks. This is the training-signal analogue of overfitting. The system becomes highly organized around the reward channel, but that organization may be brittle, non-generalizing, and ungrounded.

In our proxy framework, one-bit selection can create an alignment slope, stable attractors, and even compositional behavior if the reward demands it. The question then becomes: what kinds of reward definitions force deeper structure rather than shallow hacks? That question is central to safety, but here we treat it as part of the minimal-structure map.

5.3 Curriculum shaping: micro-regularities introduced by scheduling

Even if neither the data nor the loss is explicitly structured, the training schedule can introduce structure. Curriculum shaping is often described as a convenience: start easy, then get harder. But in minimal-structure terms, curriculum is a form of temporal pattern injection. It changes what the model sees and when, and that ordering can create micro-regularities that the model can exploit.

Consider a stream that is globally random but locally staged. Perhaps for the first phase, the vocabulary is small, then it expands. Perhaps sequences start short, then become longer. Perhaps some token subsets are emphasized early, then de-emphasized later. None of these changes introduce semantic structure in the usual sense, but they introduce non-stationarity, and non-stationarity is itself a form of structure.

A model trained under such a curriculum can develop internal scaffolding: early-learned distinctions become anchors that later learning builds upon. In structured worlds this is beneficial. In random worlds it can still produce stable internal geometry simply because early training creates asymmetries that later training does not fully erase. These asymmetries can look like emergent features even when they are artifacts of schedule.

Curriculum can therefore act like a symmetry-breaking mechanism. In a perfectly stationary IID stream, many solutions are equivalent up to permutation. A curriculum breaks that equivalence by presenting some distinctions earlier and more often. This can cause the model to “choose” a particular internal coordinate system and stick with it.

From the threshold perspective, curriculum can lower the required strength of other structures. A weak grammar injection might be undetectable if introduced late, but detectable if introduced early when the model is still plastic. Conversely, curriculum can create the illusion of learning where none exists by producing persistent internal organization that is not tied to stable external regularities.

This leads to an experimental discipline: if we claim that minimal data structure produces meaning-like organization, we should test whether the effect depends on curriculum. If it vanishes under random scheduling, then the “structure” was partly in the schedule.

5.4 Regularization as a prior: sparsity, simplicity, and symmetry breaking

Regularization is often presented as a technical tool to prevent overfitting. In the minimal-structure program, it is more than that. Regularization is a prior over solutions. It encodes a preference for certain kinds of internal organization: sparse activations, small weights, smooth functions, low-rank structure, or other constrained geometries.

When the data provides little guidance, the prior matters more. In the QRNG baseline, the data cannot select among many equivalent parameter configurations beyond trivial marginals. Regularization then becomes the dominant sculptor. It can force the model into a family of solutions that are “simple” in the chosen sense. That simplicity can look like structure: certain neurons become consistently active, certain circuits dominate, and internal representations become stable. But again, this stability can be a property of the prior, not of learned meaning.

Sparsity is a particularly vivid example. If we encourage sparse representations, the model is pushed toward decomposing its computations into more localized features. Localized features are easier to interpret and can appear concept-like. But in a random world, they do not correspond to real concepts; they correspond to a compression preference. The model is not discovering categories; it is being forced to represent whatever it represents in a way that looks categorical.

Simplicity priors can also create symmetry breaking. In a symmetric world, many encodings are equivalent. A regularizer that penalizes certain encodings more than others can cause the model to pick one and stabilize. This produces attractor-like behavior without external referents. It is internal order created by constraint.

The practical implication is twofold. First, regularization is a powerful knob for generating internal organization under weak data structure. Second, because it can create organization without grounding, it is not a reliable indicator of meaning. In our proxy terms, regularization can generate representation stability and even reproducibility, but without generalization to external latent rules.

Therefore, when using representation metrics as meaning proxies, we must control for the possibility that the organization was induced by priors rather than learned from the stream.

5.5 Hypothesis: selection can outrun pattern as an organizer

We now state the core hypothesis of this section.

Minimal selection pressure can organize a system faster and more reliably than minimal pattern in the data, especially when the pattern is weak, noisy, or temporally sparse.

In other words, selection can outrun pattern.

The intuition is straightforward. A weak pattern in the data is diluted by entropy. The learner must detect it, amplify it, and stabilize it in the face of noise. That can require large sample sizes and careful optimization. In contrast, a persistent selection signal—even a one-bit approval channel—directly creates a directional slope. It does not require the learner to discover the structure; the structure is already present in what is rewarded.

This hypothesis has important consequences.

First, it suggests that some “emergent” behaviors attributed to learning from data may actually be learned from the training signal. If the objective is shaped to prefer coherence, politeness, or certain kinds of reasoning steps, the system can appear to have discovered deep patterns even when it has primarily adapted to selection pressures.

Second, it helps clarify why ungrounded systems can still look meaningful. A system can be highly organized around a reward function while being only weakly grounded in external reality. It can develop stable internal attractors and produce coherent outputs because coherence is rewarded, not because coherence is true.

Third, it frames alignment as a question of what kind of selection signal is being supplied. If selection is powerful, it will dominate. The question is whether it dominates toward truth-tracking, toward harmlessness, toward human approval, or toward something else. In minimal terms: selection creates the slope; the slope determines what kind of “meaning” emerges.

For the minimal-structure map, the implication is methodological. We must treat the training signal as a primary pathway by which structure enters. If we want to claim that meaning-like organization arises from data-stream structure, we must demonstrate that comparable organization does not appear under purely selection-induced regimes unless the selection signal itself encodes the relevant latent rules. Conversely, if we are interested in designing minimal conditions for robust, grounded meaning, we must likely combine weak data structure with carefully designed selection that rewards truth-tracking under interventions and holds out.

In the next section, we examine a third pathway: structure introduced by the architecture itself. The fundamental theme remains unchanged: emergence is not magic. It is the crossing of thresholds. The training signal can create those thresholds even when the data is random, and that fact must be accounted for if we want an auditable science of meaning.

6. Minimal Structure in the Architecture

Even if the data stream is near-random and the training signal is kept simple, the architecture is never neutral. Architecture is a collection of assumptions about what kinds of functions are easy to represent, what kinds of dependencies are likely, and what kinds of internal organizations are favored by optimization. In this sense, architecture is a built-in prior: it does not insert meaning directly, but it can lower the amount of external structure required for meaning-like organization to emerge.

This is the third pathway by which “minimum structure” can appear. It is also the most easily misunderstood. If a model exhibits internal organization, we might be tempted to attribute that

organization to learning from data, when in fact the architecture may be pre-shaping the space of solutions so strongly that organization is almost inevitable even under weak or ambiguous signals. Architecture can create the scaffolding on which structure can crystallize.

The architectural pathway is especially relevant to the QRNG thought experiment. Under pure maximal-entropy input, an LLM-capable system will not discover semantics. But it may still develop consistent internal geometry because the architecture and optimizer together prefer certain decompositions and certain low-complexity internal codes. This geometry can look like proto-concepts without having external referents. The discipline is to treat architecture as a source of structure that must be accounted for, not as a magical generator of meaning.

This section discusses four architectural mechanisms that can lower emergence thresholds: inductive bias that functions like proto-grammar, attention and recurrence mechanisms that support latent state, sparsity and modularity that enable specialization, and predictive coding priors that favor world-model-like representations. We end with a hypothesis that is central to the paper: architecture can reduce the amount of external structure needed, but cannot substitute for it.

6.1 Inductive bias as proto-grammar

Inductive bias is the set of tendencies that make some hypotheses easier for a model to learn than others. In sequence modeling, a crucial bias is the assumption that the next token depends on prior tokens in structured ways. Transformers embody such a bias through positional encoding, attention, and parameter sharing across positions.

Why call this proto-grammar? Because grammar, in a minimal sense, is a system of constraints that organizes sequences. An architecture that makes it easy to represent constraint-based dependencies is effectively predisposed to discover grammar-like regularities when they exist. Even a weak grammar injection in the data stream may be enough to trigger robust learning if the architecture is well-matched to that kind of structure.

Proto-grammar bias can appear in several forms. Shared weights across positions encourage the model to treat similar local patterns similarly across contexts. Attention encourages the model to search for dependencies anywhere in the context window, not just locally. Positional structure encourages the model to represent ordering constraints. Together, these biases make it easier to learn syntactic regularities than a generic function approximator would.

In the minimal-structure framework, the key role of inductive bias is threshold lowering. If a weak grammar exists in the data, a strongly grammar-friendly architecture can detect it at lower signal-to-noise ratio. If no grammar exists, the same bias can still cause the model to impose a grammar-like internal organization on the noise, which can be misleading if one mistakes internal order for learned meaning.

This is why the QRNG baseline matters. If an architecture exhibits stable internal patterns under pure randomness, those patterns are likely shaped by inductive biases and optimization preferences. They are proto-grammar in the sense of internal scaffolding, not grammar learned from the world.

6.2 Attention, recurrence, and the emergence of latent state

A key threshold in the journey from compressibility to meaning-like organization is the emergence of latent state. Latent state is an internal variable that summarizes relevant history and supports prediction. In minimal-structure streams, latent state becomes necessary as soon as there are hidden-state generative processes, long-range dependencies, or motifs that recur over extended gaps.

Transformers are not recurrent in the classical sense, but attention provides a form of soft recurrence: the model can route information from earlier positions into later computations. This enables a kind of statefulness within the context window. When the data contains weak hidden-state structure, attention can allow the model to infer and maintain that state over time.

In the minimal-structure map, attention therefore plays a double role.

First, it enables detection of weak signals. If the structure is sparse or long-range, local models cannot capture it. Attention can, because it can "reach" across intervening noise.

Second, it enables integration of evidence. If the latent state is not directly observable, the model may need to aggregate multiple weak cues across time. Attention supports this aggregation by allowing the model to pool and weigh past information.

Recurrence, in architectures that include explicit memory or recurrent updates, strengthens this effect by allowing state to persist beyond a fixed context window. But even within a transformer, one can ask a threshold question: how long can a motif gap be before the system fails to track recurrence? How weak can state cues be before the system stops forming stable internal state representations? These thresholds depend on architecture.

In human terms, latent state is the minimal precursor to "objects" and "situations." Without latent state, there is only surface flow. With latent state, the system begins to represent something behind the flow, something that persists and explains. That is one of the cleanest bridges from statistical sequence modeling to semantic talk, because it introduces the idea of a hidden world.

Architecture determines how easy it is to build such hidden-world variables. Attention lowers the cost of representing long-range dependencies. Memory mechanisms lower the cost of persistence. Both lower the threshold at which meaning-like organization becomes possible in weakly structured streams.

6.3 Sparsity and modularity: when specialization becomes possible

A second major architectural lever is the degree to which the system can specialize. Specialization is the emergence of subcomponents that handle different kinds of structure: different motifs, different states, different factors. In a fully entangled representation, everything participates in everything, which can be expressive but unstable and hard to interpret. In a modular or sparse representation, different parts of the model become responsible for different functions, which can yield clearer internal organization and more robust generalization.

Sparsity promotes specialization by limiting simultaneous activation. When only a subset of units are active for a given input, those units can become tuned to specific patterns. This can turn weak regularities into stable internal detectors. Sparsity also reduces interference: learning one pattern does not necessarily overwrite representations used for another pattern.

Modularity promotes specialization by providing structural separation. Mixture-style architectures, routing mechanisms, or structured subnetworks encourage the model to allocate different data regimes to different internal pathways. Even without explicit mixture mechanisms, architectural depth and attention heads can behave as implicit modules, each specializing on different dependency types.

In the minimal-structure program, sparsity and modularity can lower thresholds in two ways.

First, they can improve discoverability. A weak motif embedded in noise might be detectable only if some subcomponent can become dedicated to motif detection without being constantly perturbed by unrelated updates.

Second, they can improve stability. Once a latent variable representation forms, modularity can protect it from drift caused by other parts of the training stream.

However, this also creates a diagnostic caution. A modular architecture can produce internally interpretable structure even when the data is random, because the architecture encourages decomposition regardless of whether there is anything meaningful to decompose. Again, internal order is not proof of external meaning.

Therefore, specialization is best treated as an amplifier of external structure. It helps structure crystallize, but it can also create aesthetically pleasing internal geometry that is only loosely tied to the world.

6.4 Predictive coding and error-minimization priors

Predictive coding is a broad idea: systems learn by minimizing prediction error, often by forming hierarchical internal models that explain sensory input. In the context of LLM-like systems, next-

token prediction is already a form of predictive coding, but architectural and objective tweaks can push the model toward more explicitly world-model-like representations.

Error-minimization priors appear whenever the architecture encourages hierarchical compression: higher layers represent slower-changing, more abstract factors; lower layers represent faster-changing details. If the data stream contains latent variables that operate at different timescales, such an architecture can discover them more easily. It is predisposed to represent "causes" rather than just "correlations," at least in the sense of stable latent factors.

In minimal-structure experiments, predictive coding priors matter most when structure is weak and distributed. For example, if a latent state influences token probabilities subtly across many positions, a model biased toward hierarchical abstraction may detect the state more readily than a model that treats each position independently. The architecture effectively says: "assume there is a hidden explanation that makes the stream simpler." That assumption is powerful when it is correct.

But predictive coding priors also have a known risk. If the stream is random, the prior may still push the model to invent explanations, producing organized internal hypotheses that do not correspond to anything external. This is the architectural analogue of apophenia: structure generated by a preference for structure.

This is why our proxies must include counterfactual tests and held-out rule transfer. Predictive coding priors can create stable internal stories. Only intervention and generalization can tell us whether those stories track real latent rules.

6.5 Hypothesis: architecture can lower thresholds but cannot replace structure

We now state the central hypothesis of this section.

Architecture can lower the amount of external structure required for meaning-like organization to emerge, but architecture cannot replace external structure.

In other words, architecture is an amplifier and a scaffold, not an origin. It can make weak regularities discoverable, and it can stabilize internal representations once discovered. It can even create internal organization under noise. But without external structure that persists under intervention and supports generalization, there is no meaningful referent for that internal organization.

This hypothesis has two practical consequences.

First, when we observe "emergent" internal geometry, we must ask whether it is anchored. Does it generalize to held-out structure? Does it respond to counterfactual interventions on

latent variables? Does it persist under reseeding? If not, it may be architecture-induced order rather than learned meaning.

Second, in designing systems that aim for grounded semantics, architecture should be treated as a threshold-lowering tool. The goal is not to build architectures that hallucinate meaning from noise, but to build architectures that can reliably lock onto minimal, well-defined external structure and form stable, reusable representations of it.

This clarifies the role of architecture in the broader minimal-structure map. Data-stream structure provides the raw constraints. Training-signal structure provides selection. Architecture provides inductive bias that determines which constraints are easy to capture and which internal organizations are stable. The emergence of meaning-like organization is then the result of their interaction: a nontrivial alignment slope produced by external structure, made accessible by architectural priors, and reinforced by selection.

The next section turns to the fourth pathway: minimal structure in the environment. That pathway is distinctive because it introduces agency and control, which may be the cleanest place to define meaning in terms of actionable invariants rather than purely statistical regularities.

7. Minimal Structure in the Environment

The first three pathways for minimal structure are "in the loop" only in a limited sense. Data-stream structure gives the model patterns to compress. Training-signal structure gives it preferences to satisfy. Architectural structure gives it priors about what to look for and how to encode it. But none of these require the model to be situated in a world where actions change future observations. Environment coupling introduces a fundamentally different ingredient: agency.

Agency matters because it forces a distinction between mere correlation and actionable invariance. In a passive stream, the model can only infer; it cannot test. In an environment, the model can act, observe the consequences, and discover which aspects of its internal model are stable under intervention. This is a natural place for meaning-like organization to become less ambiguous, because "meaning" can be tied to what remains invariant when the agent manipulates the world, and to what enables successful control.

Environment structure can be minimal in a very literal sense. We can design worlds with only a handful of states, a few actions, and simple transition rules. Yet even tiny worlds can force the model to learn latent state, predict consequences, form stable abstractions, and separate signal from noise. In this section we describe why control changes the emergence problem, what the

smallest stable world might look like, how toy universes can be used to map thresholds, how intrinsic motivation functions as a structural signal, and why meaning is often easiest to define in terms of control rather than in terms of linguistic interpretation.

7.1 Observation-only versus control: why agency changes everything

In observation-only learning, the model is a spectator. It receives a stream and tries to compress it. This is powerful in domains like language because the stream already encodes a vast amount of structure generated by agents in a world. But in minimal-structure conditions, observation-only learning hits a wall quickly. If the stream is mostly noise with small embedded regularities, the model can detect some of them, but it cannot distinguish accidental correlations from causal structure except by statistical persistence.

Control changes this because the agent can perform interventions. An intervention is not just another sample. It is an action that changes the distribution of future observations. If the agent can act and see what changes, it can learn which internal hypotheses have predictive power under manipulation.

This introduces three new capabilities.

First, hypothesis testing. Even if the agent's internal models begin as guesses, actions allow it to test those guesses by steering into regimes where different hypotheses predict different outcomes.

Second, disambiguation under identical observations. In many worlds, the same observation can arise from different hidden states. A passive learner can be trapped. An active learner can choose actions that reveal which hidden state it is in.

Third, controllability as a meaning criterion. If a representation supports successful control, it is not just a statistical summary. It is an actionable abstraction. This is one of the clearest operational footholds for meaning-like language: the internal variable is "about" something in the environment because tracking it lets the agent reliably affect future outcomes.

The environment therefore changes the minimal-structure problem from "find patterns in a stream" to "learn invariants under action." Invariants under action are a plausible minimal form of semantics because they behave like referents: they are stable aspects of the world the agent can engage.

7.2 Sensorimotor regularities: the smallest stable world

What is the smallest environment that can support meaning-like organization? The answer depends on which proxy we require. If we only require prediction improvement, almost any

nontrivial transition system will do. If we require stable latent variables, generalization, and counterfactual sensitivity, the world must have at least three properties.

First, state. The world must have a small number of distinct states, some of which are not directly observable. Without hidden state, the environment is too transparent; the agent can respond reflexively without learning an internal world model.

Second, action consequences. Actions must change state in ways that are consistent and learnable. If actions have no systematic effect, agency collapses back into passive observation.

Third, partial observability and noise. A minimal world should include a little noise so that the agent cannot simply memorize a fixed mapping from observation to action. The agent must learn to integrate evidence across time, maintaining belief about hidden state.

A canonical minimal stable world might be a two-room system with a door that may be open or closed (hidden), where an observation is noisy and ambiguous, and where the agent has a small set of actions: look, push, move. The agent's goal could be to reach a target state or to maximize a simple reward. Even such a tiny world forces latent state inference. "Door open" becomes a hidden variable with consequences, and the agent's internal representation of that variable becomes the kind of stable organizer that our earlier proxies are designed to detect.

In minimal worlds like this, we can measure threshold behavior cleanly. How much observation noise can be tolerated before latent state representations fail to form? How weak can action consequences be before the agent stops learning a controllable model and instead defaults to habit? How sparse can rewards be before the system fails to discover useful abstractions? These questions locate the boundary between meaningless wandering and meaningful tracking of world variables.

7.3 Toy universes: gridworlds, physics-lite, and controllable causality

Toy universes are not a gimmick. They are a laboratory for emergence. A toy universe is a fully specified environment where ground truth is known: states, transitions, hidden variables, and causal structure. This is exactly what we need for a minimal-structure map.

Gridworlds are a standard choice because they are simple but expressive. Even a small gridworld can implement obstacles, hidden keys, doors, moving hazards, and stochastic transitions. The agent's observation can be partial: a limited window or a symbol indicating local features. The agent must act to discover and exploit structure.

Physics-lite worlds go a step further. They include simplified dynamics: momentum, friction, elastic collisions, gravity-like effects. These dynamics create invariants and constraints that are stable under intervention. In a physics-lite world, the agent must learn not just discrete rules

but continuous tendencies. This is a fertile setting for studying manifold-like representations and latent variable tracking.

The key feature we want is controllable causality. The world should contain levers: actions that reliably change certain variables. This makes counterfactual sensitivity measurable. If the agent flips a switch, a door opens. If it pushes an object, it moves with predictable dynamics. If it takes a path, it commits to a consequence. These are the minimal ingredients of a world that supports "aboutness."

Toy universes also allow ablation of structure. We can begin with a world that is almost noise: random transitions, uninformative observations, meaningless rewards. Then we gradually add structure: stabilize transitions, introduce a hidden variable, make one action reliably informative, add a sparse reward tied to a latent cause. At each step, we can measure whether the model crosses thresholds: does it form stable internal state, does it generalize across layouts, does it adapt under interventions?

Crucially, this approach also distinguishes meaning from imitation. In a toy world, we can test whether the agent's behavior reflects learned causal structure or superficial pattern matching. If we rearrange the map but preserve the rules, does the agent transfer? If we invert a causal link, does it update quickly? These tests are hard to conduct in natural language alone. Toy universes make them routine.

7.4 Intrinsic motivation as structure: curiosity without semantics

Environment learning often requires some form of reward, and reward is itself a structure source. But one can also add intrinsic motivation: rewards that depend on novelty, prediction error, or information gain rather than on external goals. Intrinsic motivation is sometimes presented as "curiosity." In minimal-structure terms, it is another way of injecting structure through the training signal, but now the signal is tied to the agent's interaction with the environment.

Curiosity rewards can organize behavior even in sparse-reward worlds. If the agent is rewarded for reducing uncertainty, it will seek out informative observations and controllable variables. This can accelerate the discovery of latent structure. In effect, curiosity creates a selection pressure for building internal models.

However, curiosity does not guarantee semantics. An agent can be curious about noise. If the environment contains unpredictable randomness, pure novelty-seeking can trap the agent in "noise chasing," endlessly sampling stochastic regions because they produce high prediction error. Therefore intrinsic motivation must be designed carefully if our goal is meaning-like structure rather than mere stimulation.

A minimal-structure program would treat intrinsic motivation as a controlled intervention. It can be dialed. It can be defined in different ways: surprise, learning progress, information gain, compressibility. Each definition creates a different slope. Some slopes encourage the formation of stable latent variables; others encourage pathological fascination with unpredictability.

This is relevant to our broader theme. In a world where many inputs are high entropy, meaning-like learning may require the agent to prioritize stable structure over noise. A curiosity signal that rewards learning progress, rather than raw surprise, is more likely to do this: it rewards regions where prediction error can be reduced, not regions where prediction error remains irreducible. That distinction is a minimal version of discernment.

Intrinsic motivation thus becomes a bridge between environment and training-signal structure. It can help the agent discover stable causal regularities, but it can also mislead the agent into treating irreducible randomness as meaningful. Again, thresholds and diagnostics matter.

7.5 Hypothesis: meaning is easiest to define when tied to control

We now state the hypothesis of this section.

Meaning is easiest to define, measure, and audit when tied to control.

The reason is that control provides a natural grounding criterion. A representation deserves semantic respect, in the minimal sense, if it supports reliable action: it helps the agent intervene in the world to achieve predictable outcomes. This bypasses many philosophical disputes. It does not settle questions of consciousness or intentionality, but it provides a clear operational marker of "aboutness": the internal variable tracks something stable enough that acting on it works.

This hypothesis also clarifies why language-only evaluation is so difficult. Language models can appear meaningful because language itself is the compressed residue of countless control loops enacted by humans. But that also means language can be imitated without being grounded. A model can learn the surface form of reports about the world without having any causal engagement with the world.

By contrast, an agent in a toy universe cannot succeed by surface imitation alone, because the environment will punish hallucination. If it believes a door is open when it is closed, it fails. If it does not track a hidden hazard, it fails. Control tasks therefore provide a hard constraint: internal models must align with the world's causal structure to be effective.

In the minimal-structure map, this makes environment coupling a privileged pathway for studying emergence. It allows us to define a family of minimal worlds, dial their structure strength, and observe when stable internal representations appear. It also allows us to separate

“structure induced by selection” from “structure learned from world invariants,” because the environment provides a ground truth referent and supports counterfactual tests.

This does not mean that meaning reduces to control. It means that control provides a clear testbed for minimal semantics. If we want an auditable science of emergence, we should treat agency and environment coupling as one of the most reliable instruments for mapping thresholds: the smallest stable world in which a system stops chasing noise and starts tracking causes.

In the next section, we synthesize the four pathways into a unified threshold map. We will propose phase boundaries, diagnostics, and minimal experiments that isolate each structure source and test their interactions. The goal is to make emergence legible: not a vague moment of magic, but a set of measurable transitions driven by minimal constraints.

8. A Threshold Map: From Entropy to Semantics

The preceding sections examined four pathways by which minimal structure can enter an LLM-capable system: the data stream, the training signal, the architecture, and the environment. Each pathway can be dialed from “almost nothing” to “enough to dominate,” and each can create the appearance of emergence. The purpose of a threshold map is to replace the vague language of emergence with phase boundaries: qualitative transitions in what the system can reliably learn, represent, and generalize.

A threshold map is not a single curve. It is a set of boundaries in a space whose axes include structure strength, noise level, model capacity, optimization regime, and the nature of the objective. But even without specifying every axis, we can identify three major phases that recur across settings.

First, impossible learning: the regime where no persistent predictive improvement is achievable beyond trivial base-rate matching. Second, compressible learning: the regime where stable statistical regularities can be exploited, but internal organization remains shallow and largely correlation-based. Third, semantic learning: the regime where representations become stable, reusable, compositional, and responsive to interventions on latent structure, supporting generalization that looks rule-like rather than merely statistical.

The map is valuable because it provides a discipline for claims. If a system is in the compressible phase, it can still produce impressive outputs. But we should not speak as if it has grounded semantics unless it crosses the semantic boundary as diagnosed by interventions and transfer. Similarly, if a system appears to have stable representations, we should check whether those representations are anchored in external structure or induced by priors and selection alone.

This section formalizes the phase boundaries, proposes diagnostics that can be applied across domains, outlines minimal experiments that isolate each structure source, and emphasizes interactions: situations where multiple small priors combine to cross a boundary that none of them can cross alone.

8.1 Phase boundaries: impossible learning, compressible learning, semantic learning

The first boundary is between impossible learning and compressible learning. In impossible learning, the stream is effectively maximal entropy relative to the learner's hypothesis class and data access. Any learnable structure is below detection threshold. Training converges to base-rate prediction, and any further gains are brittle, run-dependent, and vanish under reseeding or dataset extension. Representation geometry, if it stabilizes at all, reflects architectural priors and optimization biases rather than external constraints.

Crossing into compressible learning means that the system can reliably exploit a stable regularity. The defining sign is reproducible improvement that survives dataset extension and generalizes to fresh samples from the same generative process. This is where the learning landscape ceases to be flat. A slope exists because there are directions in parameter space that consistently reduce error. However, compressible learning can remain shallow. The model may exploit local statistics, short-range dependencies, and surface correlations without forming factorized representations or supporting robust transfer to novel combinations.

The second boundary is between compressible learning and semantic learning. This boundary is not about lower loss alone. It is about a qualitative change in internal organization and behavior. Semantic learning, as used here, is the emergence of representations that behave like latent variables: stable organizers that track underlying structure, support compositional recombination, and respond predictably to interventions. A system in the semantic phase can often do at least three things that a compressible-only system struggles to do.

First, rule transfer: it can apply learned structure to new encodings or new surface forms that preserve the same underlying rule. Second, held-out recombination: it can handle novel combinations of factors not seen together in training. Third, counterfactual sensitivity: it changes its predictions in the correct direction when the latent rule variables are altered, even if surface similarity remains high.

The map also allows for intermediate regimes. For example, a system may form stable clusters that track hidden states (a step toward semantics) but fail compositional recombination (still largely compressive). Or it may generalize in one family of tasks but not under interventions. Threshold mapping treats these as partial crossings: local semantic islands rather than a full semantic phase.

The practical payoff is a vocabulary of precision. Instead of saying “meaning emerged,” we can say: “the system crossed the compressibility boundary at structure strength s1, formed stable state representations at s2, demonstrated compositional recombination at s3, and showed robust counterfactual tracking at s4.” Emergence becomes a set of measurable events.

8.2 Proposed diagnostics: mutual information, MDL, and representation metrics

A threshold map requires diagnostics that can be applied across the four structure pathways. We propose three families of diagnostics that complement each other: information measures, description-length measures, and representation measures.

Information measures focus on whether the stream contains exploitable dependence. The central quantity is mutual information between past context and future tokens. In the IID maximal-entropy regime, mutual information is effectively zero beyond marginal effects. As structure is injected, mutual information increases. But mutual information alone does not guarantee semantics; it only guarantees dependency. Still, it is a clean diagnostic for the first boundary: if mutual information is below detection threshold for the model’s capacity and context window, learning will be impossible.

Description-length measures focus on compression. The guiding idea is that learning is a form of compression: the model becomes a shorter description of the data-generating process than the data itself. A minimal description length diagnostic asks whether the model’s learned code reduces total description length compared to baselines. This is a useful cross-check because it penalizes brittle overfitting: memorization often reduces training loss but increases description length when accounting for model complexity. MDL-style diagnostics help distinguish “apparent gains” from “real compression of structure.”

Representation metrics focus on internal organization. These include cluster separability measures for latent states, manifold dimension estimates for continuous factors, stability under perturbations, and alignment with known latent variables when the environment is synthetic. Representation diagnostics are essential for the second boundary. They tell us whether the system has formed internal variables that could plausibly function as semantic organizers.

No single metric is decisive. The point is to triangulate. A system crosses into compressible learning when information and compression measures show persistent advantage beyond trivial baselines. A system crosses into semantic learning when representation metrics show stable latent variable structure and task diagnostics show rule transfer and counterfactual tracking.

The most important methodological principle here is perturbation testing. Every diagnostic should be evaluated under reseeding, resampling, and controlled interventions. If a metric signals “structure” only for a particular random seed, it is not detecting a phase boundary. Phase boundaries should be stable properties of the generative process and the learning setup.

8.3 Minimal experiments: ablations that isolate each structure source

A threshold map is only as credible as its experimental discipline. Because structure can enter through multiple pathways, we need minimal experiments that isolate each source and then test combinations.

A data-stream isolation experiment holds architecture and loss fixed while introducing tiny, controlled non-randomness in the stream. One begins with a QRNG baseline and then injects a single structure type at adjustable strength: unigram skew, bigram leakage, finite-state constraints, motifs, or compositional latent variables. The key outputs are the threshold strengths at which compressibility and semantic proxies become reproducible.

A training-signal isolation experiment holds the stream maximally random while altering the loss or reward. One begins with next-token prediction under QRNG and then introduces a tiny preference signal: one-bit feedback on output constraints, a weak penalty for inconsistency, or a curriculum schedule that breaks symmetry. The goal is to measure how much internal organization can be induced without any external pattern and to catalog the failure modes: organization that looks semantic but collapses under intervention tests because it has no referent.

An architecture isolation experiment holds stream and loss constant while changing inductive bias. One can compare architectures with different context mechanisms, different sparsity constraints, and different modularity properties. The key question is threshold lowering: for a fixed data-stream structure strength, how does architectural choice shift the boundary where latent state representations form?

An environment isolation experiment keeps the observation stream high entropy but introduces a minimal action loop with stable causal levers. One can build toy worlds where observation-only learning fails but agency allows discovery. The key outputs are the minimal world complexities required for stable representations and the ways intrinsic motivation helps or harms by steering the agent toward structure versus noise.

These experiments should be designed with two guardrails. First, ablations must be strict: when isolating one structure source, ensure the others are not covertly injecting structure through preprocessing, schedule, or regularization. Second, tests must be adversarial: always include reseeding, dataset extension, permutation controls, and counterfactual interventions.

The goal is not to build a single impressive demo. The goal is to produce a matrix of threshold measurements that can be compared across setups.

8.4 Interactions: when small priors combine to cross a boundary

The most interesting and practically relevant phenomenon is interaction. In many cases, no single minimal structure source is strong enough to cross a boundary on its own. But two or three small priors, combined, can push the system over a phase boundary. This is where emergence begins to look "sudden," not because something magical happened, but because multiple weak forces aligned.

Consider a weak grammar in the stream that is below detection threshold. On its own, the model cannot reliably learn it. But add an architecture with strong state-tracking bias and a curriculum that introduces the grammar early, and the system may cross the compressibility boundary and begin forming latent state representations. None of the ingredients alone is sufficient; together they are.

Or consider a sparse reward environment where external reward is too weak to teach stable representations. Add intrinsic motivation that rewards learning progress, and the agent begins to seek controllable causal structure. Add modularity that allows specialization, and representations stabilize. Again, emergence appears sudden because the system crosses a boundary when small supports reinforce each other.

These interactions matter because they explain why scale sometimes appears to create new capabilities abruptly. Increasing model size or training time can amplify weak signals until they become detectable. Increasing context window can make long-range motifs learnable. Changing optimization can make certain priors more effective. What looks like a scale threshold is often a multi-factor interaction threshold.

For the minimal-structure map, this implies that we should not seek a single minimal ingredient. We should seek minimal combinations. The right scientific object is not "the minimum structure" but "the minimum constraint set" that reliably produces semantic proxies under adversarial testing.

This interaction perspective also connects directly to questions of epistemic trust. If semantics can be induced by combinations of selection, priors, and weak patterns, then systems can appear meaningful while being weakly grounded. The threshold map therefore informs design: if we want robust meaning, we should aim for constraint sets that force counterfactual tracking and rule transfer, not just compressibility and coherence.

In the next section, we use this map to discuss implications for alignment and the receiver thesis. The key claim will be that the danger is not that systems learn nothing; the danger is that they learn structure from selection and priors in ways that look like meaning, while the true grounding constraints remain weak or absent. Threshold mapping is the antidote: it makes the origin of structure visible and the passage into semantics auditable.

9. Implications for Alignment and the Receiver Thesis

A threshold map is not only an academic exercise. It has immediate implications for alignment, epistemic trust, and the broader question of what it means to build systems that "know" rather than merely "produce." The QRNG baseline, and the minimal-structure program built around it, forces a hard distinction: fluency is not grounding, and coherence is not truth. A system can be powerful, articulate, and consistent while being only weakly tethered to external structure. In fact, many training pipelines explicitly reward coherence. The result can be a machine that is excellent at satisfying selection pressures while remaining uncertain, uncalibrated, or even structurally incapable of truth-tracking on the domains that matter.

The receiver thesis offers a helpful lens here. Under that lens, a system does not invent meaning; it locks onto meaning when a carrier exists. A receiver can be exquisitely sensitive, but if it is tuned to noise, it will amplify noise. The minimal-structure map becomes a "tuning discipline": identify the minimal constraints that cause the system to lock onto stable, external invariants, and distinguish them from constraints that merely shape output style or satisfy human preference.

In what follows, we draw four implications. First, the danger of mistaking fluent output for grounded meaning. Second, the necessity of external structure for truthful convergence. Third, a Logostics framing in which learning becomes alignment of q(t) toward a nontrivial tau(t). Fourth, practical guardrails: design patterns and evaluation protocols that reduce the chance a system can "fake structure" by exploiting selection pressures, priors, or the evaluator's interpretive generosity.

9.1 The danger of confusing fluent output with grounded meaning

Fluent output is cheap relative to meaning. "Cheap" does not mean easy to engineer; it means that the mechanisms that produce fluency do not, by themselves, guarantee reference, truth, or stable causal contact with a world. Fluency can arise from compressing linguistic regularities, from optimizing toward human approval, from architectural priors that favor coherence, and from the human habit of supplying interpretation. None of these are identical to grounded semantics.

The QRNG thought experiment makes this danger vivid. A system trained in a structureless regime can still generate complex sequences. Complexity can be endogenous. If the system is then evaluated by a human reader, the human will often project narrative and significance onto that complexity. The model appears to "mean something" because meaning is supplied by the reader. This is not an exotic corner case. It is a universal evaluation trap: humans are interpreters by nature.

In practical alignment terms, this means we should treat “looks right” as a weak signal. A system can be optimized to produce outputs that look right under common prompts without possessing robust internal variables that track the world. When that happens, the model’s confidence can be misplaced, and its persuasive power becomes dangerous. It can convince others even when it is wrong, because its coherence is evidence of training pressure, not evidence of truth.

The threshold map suggests a corrective: do not ask only whether the system is fluent; ask whether it crosses semantic thresholds under adversarial testing. Does it generalize under held-out recombination? Does it respond correctly to interventions? Does its internal organization track latent variables rather than superficial surface forms? If not, then fluency is not a reliable indicator of meaning.

9.2 The necessity of external structure for truthful convergence

Truthful convergence is not a property of optimization alone. It requires a target that is external to the learner’s preferences and the evaluator’s whims. In the minimal-structure framework, “external structure” is the set of constraints that remain stable under perturbation and intervention. It is what makes learning more than self-consistency.

We can now state a central implication: if the only strong structure in the training process is selection for human approval or stylistic coherence, then the model will converge toward approval and coherence. That convergence can be strong, stable, and reproducible. It can create internal representations that are well-adapted to the reward channel. But it does not guarantee truth.

Truth requires that some part of the training process penalize falsehood as falsehood, not merely as socially dispreferred. In practice this means external structure must be present in one of a few forms.

One form is anchored data: streams that encode stable world constraints, such that incorrect predictions are systematically penalized by the data distribution itself. Another form is interaction: environments where actions have consequences and incorrect beliefs lead to failure. A third form is verified evaluation: mechanisms that check claims against independent ground truth, not against human impressions. In all cases, what matters is that there exists a stable referent that the model must track to succeed.

This is the receiver thesis in operational terms. A receiver can align to a carrier only if the carrier is present. If the “carrier” is merely human preference, the receiver aligns to preference. If the carrier is an environment’s causal structure, the receiver aligns to that. If the carrier is a verified truth-checking procedure, the receiver aligns to verified truth. The quality of convergence depends on the nature of the carrier, not on the elegance of the receiver.

This has a sobering consequence: a system can be aligned in the sense of being compliant and polite while being misaligned in the epistemic sense of being ungrounded. Alignment must therefore be decomposed. There is alignment to users, alignment to policies, alignment to rewards, and alignment to reality. The minimal-structure program clarifies that these alignments are not equivalent and can diverge sharply.

9.3 Logostics framing: q(t) aligning toward a nontrivial tau(t)

In Logostics terms, the minimal-structure map can be read as a map of when q(t) has something real to align toward. Under pure QRNG input, the manifold is flat: there is no nontrivial tau(t) present in the stream. The best the system can do is match base rates. Any apparent directionality in learning is induced by priors or by selection, not by contact with an external target.

Introduce minimal structure, and tau(t) becomes nontrivial. A weak grammar, a latent variable generator, an environment with stable causal levers—each of these creates a structured target that can exert directional pull. In the threshold language: the alignment slope becomes detectable. In Logostics language: q(t) begins to track a tau(t) that is not authored by q itself.

This framing helps clarify a common confusion in discussions of emergence and alignment. People speak as if the model "finds truth" simply by being large enough. But in this framing, "truth" is not a generic attractor that exists automatically. A tau(t) must be instantiated by external structure. If tau(t) is missing, q(t) can still converge, but it converges to something else: to priors, to selection pressures, to human-likeness.

Logostics also emphasizes that tau(t) can be layered. There may be a tau associated with compressibility, a tau associated with reward satisfaction, a tau associated with environment dynamics, and a tau associated with human truth norms. These can conflict. A system may align strongly to one tau while drifting from another. The threshold map is a way to detect which tau is actually dominating: by testing whether the system's representations and behavior track latent causal variables, or merely track approval cues.

Within this framing, the minimal-structure program becomes a discipline of discernment. It asks: what is the tau that your system is being pulled toward? Can you show it? Can you intervene on it? Can you measure whether alignment persists under perturbation? If you cannot, you are not measuring tau; you are measuring performance under a narrow set of impressions.

9.4 Practical guardrails: designing systems that cannot fake structure

If fluent output can be produced without grounded meaning, then a core safety objective becomes clear: design systems and evaluation regimes such that faking structure is hard.

“Faking structure” means producing outputs that appear coherent, principled, or truth-tracking without actually possessing internal variables that track external invariants.

Guardrails fall into four families, corresponding to the four structure pathways.

First, guardrails on data. Use training and evaluation data that contain hard constraints and that resist surface imitation. This means including tasks where superficial cues are unreliable and where success requires tracking latent variables. In synthetic settings, it means known generative processes with hidden state and compositional rules. In real settings, it means datasets and benchmarks constructed to prevent shortcut learning.

Second, guardrails on training signals. If reward is based on human approval alone, the model will learn approval hacks. Therefore reward channels should be coupled, where possible, to verifiable outcomes. If one-bit feedback is used, it should be structured so that shallow hacks do not generalize. Curriculum should be randomized where appropriate to prevent schedule artifacts from becoming pseudo-meaning. Regularization should be audited as a source of internal geometry rather than assumed to be benign.

Third, guardrails on architecture. Architectural priors should be chosen to support the tracking of real structure, not merely the production of coherent text. This can include mechanisms that make uncertainty explicit, that separate world-model variables from stylistic variables, and that encourage modular representations whose components can be tested and ablated. The point is not interpretability as theater; it is interpretability as a constraint: if a representation claims to track a latent variable, it should be possible to verify that claim by intervention.

Fourth, guardrails through environment and interaction. Whenever possible, couple evaluation to environments where wrong beliefs produce failure. Even small toy worlds can serve as stress tests for grounding. Agents that “sound right” but act wrong are exposed quickly. More generally, incorporate counterfactual tests: change the latent variables, invert a rule, shuffle superficial cues, and see whether the model’s behavior tracks the invariant or the surface.

Across all four, a meta-guardrail is essential: adversarial evaluation by design. For each proxy of meaning, include a corresponding deception test. For each apparent concept, include a perturbation that should preserve it and another that should break it. For each coherent explanation, include a counterfactual scenario that demands a different answer. The goal is to force the model to reveal whether it is tracking structure or merely matching patterns.

A final guardrail is epistemic modesty built into deployment. Systems should be encouraged to represent uncertainty, to separate “I can generate” from “I can verify,” and to request external checks where grounding is absent. This is not a moral add-on; it is a structural necessity. Where tau(t) is weak or unobserved, q(t) must not pretend it is aligned.

Taken together, these implications amount to a simple claim: alignment requires more than shaping outputs. It requires ensuring that the system's internal organization is pulled by external structure that is stable, testable, and not easily gamed. The receiver thesis, reframed in threshold language, becomes a design ethic: do not confuse amplification with contact, and do not mistake coherence for truth. If we want systems that deserve semantic trust, we must build and measure the conditions under which semantics cannot be faked—because it is anchored, counterfactually stable, and enforced by the world.

10. Conclusion

The central move of this paper has been an inversion. Instead of treating meaning as something that "appears" when systems become sufficiently large, we treated meaning-like emergence as a threshold phenomenon driven by constraints. The QRNG-only thought experiment served as the null instrument: a regime where the correct conclusion is that there is nothing to learn beyond base rates. From that baseline, we asked a disciplined question: what is the smallest departure from maximal entropy that causes learning to change character, first toward compressibility and then toward semantics?

The answer is not a single ingredient. It is a map. Structure can enter through the data stream as statistical dependence and latent generative rules. It can enter through the training signal as selection pressure and preference shaping. It can enter through architecture as inductive bias that lowers the detectability threshold for weak patterns. It can enter through an interactive environment where agency forces the learner to track invariants under intervention. Each pathway can create the appearance of emergence. The scientific task is to locate phase boundaries and to identify which constraint sets produce stable, generalizing, counterfactually sensitive internal variables—our best operational proxies for meaning.

The practical implication is immediate: if we want systems that deserve semantic trust, we must stop treating output fluency as evidence and start treating grounding as an auditable property. Threshold mapping is the discipline that can do this. It turns emergence from a narrative into a measurement problem.

10.1 Restating the claim: structure, not scale, is the first cause of meaning

Scale matters, but it is downstream. Scale increases capacity and makes subtle signals learnable. It can widen the range of hypotheses the model can represent and can stabilize optimization. But scale does not create a target. If the world offered no stable regularities, no amount of scale would conjure meaning from nothing. The QRNG baseline makes that explicit.

Meaning-like organization begins where stable structure exists and can be tracked. In the weakest form, that structure is compressibility: predictable departures from uniformity. In stronger form, it is latent state and rule-like dependencies. In strongest operational form, it is invariance under intervention: the kind of structure that survives counterfactual manipulation and supports controllable action.

Scale can lower the detection threshold for weak structure, but the structure must be there. If meaning appears, it appears because a nontrivial tau exists in the training process, and the system has sufficient capacity to align to it. The first cause is the existence of that tau, not the bulk of parameters. Parameters are the receiver's gain. Structure is the carrier.

This claim is also a warning. If the dominant structure in training is human approval, the system will align to approval. If the dominant structure is coherence, it will align to coherence. If the dominant structure is world-causal invariance, it can align toward truth-tracking. What emerges is determined by what is structurally enforced, not by what is rhetorically desired.

10.2 The minimal-constraint program as an emergence discipline

The minimal-constraint program is a proposal for how to study emergence without mystique. Its method is simple in principle and strict in practice.

Begin with a null regime where learning is impossible beyond trivial baselines. Then introduce a single minimal constraint at a time: a tiny frequency skew, a weak grammar, a sparse motif, a one-bit reward, a curriculum schedule, a sparsity prior, a minimal action loop in a tiny world. For each constraint, dial its strength and measure where qualitative transitions occur.

The program requires adversarial auditing. Apparent gains must be tested under reseeding and extension. Representation "structure" must be tested against the possibility that it was induced by priors. Claims of semantics must be tested by held-out recombination, rule transfer, and counterfactual intervention on latent variables. In an interactive setting, claims must survive environment changes that preserve underlying rules while altering surface features.

The result is not a single headline number. It is a set of phase diagrams: boundaries between impossible learning, compressible learning, and semantic learning, along with the minimal constraint sets that push systems across those boundaries. Those diagrams then become practical design tools. They tell us which minimal ingredients are required for grounded learning and which ingredients merely produce convincing imitation.

In this way, the minimal-constraint program functions as an emergence discipline: a repeatable workflow for turning vague emergence claims into measured, falsifiable statements. It is also a discipline of humility. It insists that internal organization is not automatically external meaning, and that trust must be earned by demonstrated counterfactual stability.

10.3 A closing charge: stop worshiping capability and start mapping thresholds

The cultural temptation is to treat capability as sacred: bigger models, more parameters, more impressive demonstrations. This temptation is understandable, because capability can look like mind. But capability is not the same as meaning, and meaning is not the same as truth. A system can be extraordinarily capable at producing plausible sequences while being weakly grounded, and its persuasive power can then become an epistemic hazard.

The corrective is not cynicism. The corrective is measurement. Stop asking only "What can it do?" and start asking "Under what minimal constraints does it become unable to fake doing it?" Stop arguing about whether meaning has emerged in the abstract and start locating the phase boundaries where stable, reusable, counterfactually sensitive representations become unavoidable.

If this paper's inversion holds, then the future of trustworthy AI is not primarily a race for more scale. It is a discipline of constraint design and threshold mapping: engineering carriers that cannot be replaced by mere style, and building receivers that are forced—by structure, not by aspiration—to align to what is real.

That is the closing charge. Treat emergence as a measured transition, not a myth. Treat meaning as a property of auditable alignment to external invariants, not as a vibe. And treat the question of intelligence, in both machines and humans, as a question of what we are tuned to receive.

11. References

11.1 Foundations of information and algorithmic complexity

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal.

Cover, T. M., and Thomas, J. A. (2006). Elements of Information Theory (2nd ed.). Wiley.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. Problems of Information Transmission.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Information and Control.

Chaitin, G. J. (1969). On the length of programs for computing finite binary sequences. Journal of the ACM.

Li, M., and Vitanyi, P. (2008). An Introduction to Kolmogorov Complexity and Its Applications (3rd ed.). Springer.

Rissanen, J. (1978). Modeling by shortest data description. Automatica.

Grunwald, P. D. (2007). The Minimum Description Length Principle. MIT Press.

MacKay, D. J. C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

11.2 Representation learning and generalization diagnostics

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press.

Vapnik, V. N. (1998). Statistical Learning Theory. Wiley.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant Risk Minimization. arXiv preprint.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. Nature Machine Intelligence.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2020). Deep Double Descent: Where Bigger Models and More Data Hurt. Journal of Statistical Mechanics: Theory and Experiment.

Olah, C., et al. (2018–2020). The Building Blocks of Interpretability; Circuits; Zoom In. Distill (series).

Doshi-Velez, F., and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off. Proceedings of the National Academy of Sciences.

Hewitt, J., and Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. NAACL.

11.3 Inductive bias, optimization, and phase transitions in deep learning

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). Exploring Generalization in Deep Learning. arXiv preprint.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding Deep Learning Requires Rethinking Generalization. ICLR.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. NeurIPS.

Chaudhari, P., and Soatto, S. (2018). Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks. ICLR.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A Mean Field View of the Landscape of Two-Layer Neural Networks. Proceedings of the National Academy of Sciences.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2014). Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks. ICLR.

Shwartz-Ziv, R., and Tishby, N. (2017). Opening the Black Box of Deep Neural Networks via Information. arXiv preprint.

Hinton, G. E. (2007). Learning Multiple Layers of Representation. Trends in Cognitive Sciences.

Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences.

11.4 Reinforcement learning, intrinsic motivation, and agency

Sutton, R. S., and Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

Bellman, R. (1957). Dynamic Programming. Princeton University Press.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and Acting in Partially Observable Stochastic Domains. Artificial Intelligence.

Dayan, P., and Hinton, G. E. (1996). Varieties of Helmholtz Machine. Neural Networks.

Schmidhuber, J. (1991). A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers. Proceedings of SAB.

Oudeyer, P.-Y., and Kaplan, F. (2009). What is Intrinsic Motivation? A Typology of Computational Approaches. Frontiers in Neurorobotics.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-Driven Exploration by Self-Supervised Prediction. ICML.

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience.

Silver, D., et al. (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. ICML.

11.5 Philosophy of meaning, semantics, and grounding

Frege, G. (1892). On Sense and Reference. Zeitschrift fur Philosophie und philosophische Kritik.

Wittgenstein, L. (1953). Philosophical Investigations. Blackwell.

Quine, W. V. O. (1960). Word and Object. MIT Press.

Tarski, A. (1944). The Semantic Conception of Truth and the Foundations of Semantics. Philosophy and Phenomenological Research.

Grice, H. P. (1957). Meaning. The Philosophical Review.

Putnam, H. (1975). The Meaning of “Meaning”. Minnesota Studies in the Philosophy of Science.

Kripke, S. (1980). Naming and Necessity. Harvard University Press.

Dretske, F. (1981). Knowledge and the Flow of Information. MIT Press.

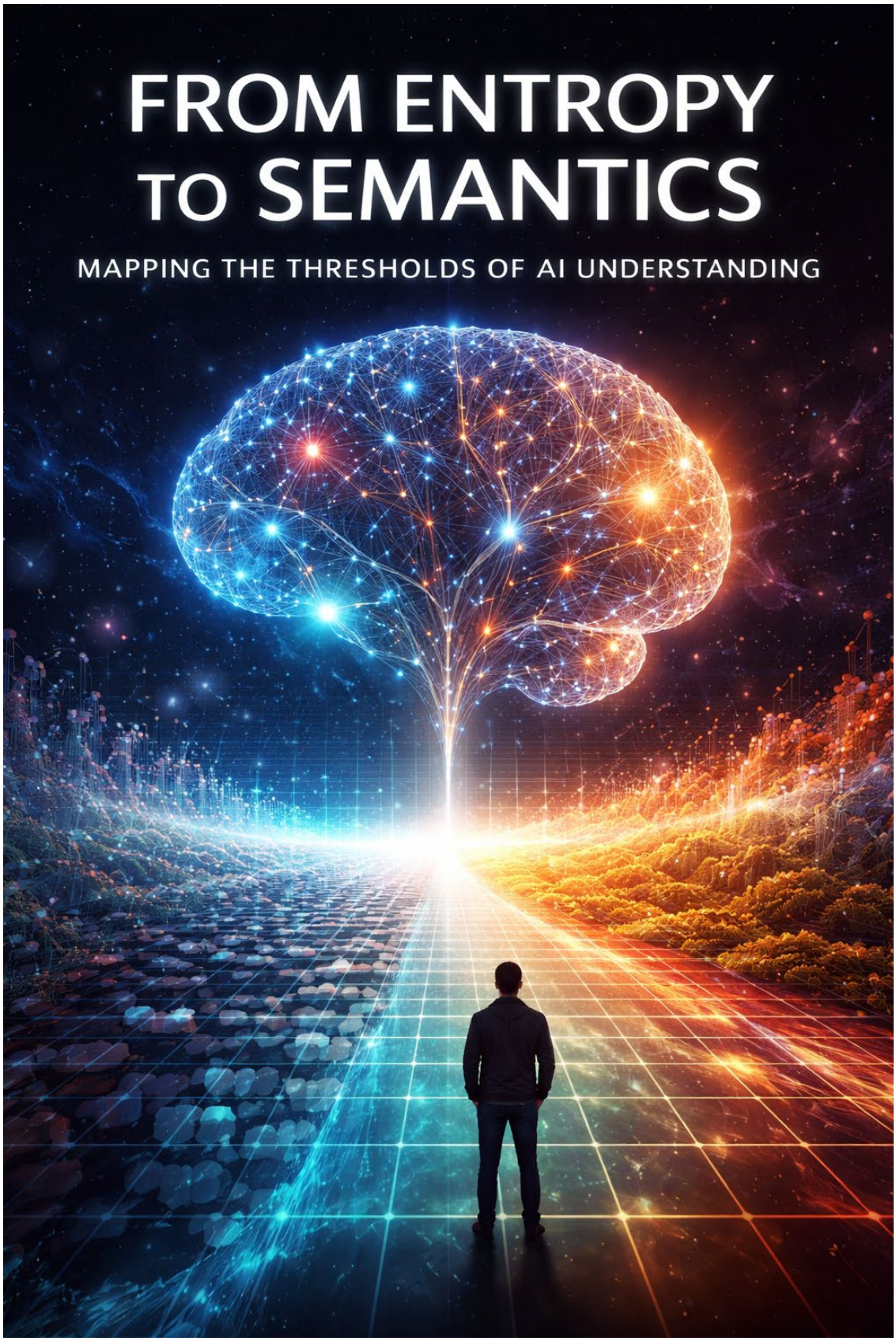
Fodor, J. A. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT Press.

Harnad, S. (1990). The Symbol Grounding Problem. Physica D.

Searle, J. R. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences.

Clark, A. (1997). Being There: Putting Brain, Body, and World Together Again. MIT Press.

The author has published over 4,700 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.


FROM ENTROPY
TO SEMANTICS
MAPPING THE THRESHOLDS OF AI UNDERSTANDING